

October 25, 2007
SC705: Advanced Statistics
Instructor: Natasha Sarkisian
Class notes: Introduction to Structural Equation Modeling (SEM)

SEM is a family of statistical techniques which builds upon multiple regression, and incorporates and integrates path analysis and factor analysis.

Advantages of SEM compared to multiple regression:

- more flexible assumptions (particularly helpful to deal with multicollinearity)
- use of confirmatory factor analysis to explicitly account for measurement error by having multiple indicators per latent variable
- graphical modeling interface (diagrams)
- ability to test models overall rather than coefficients individually
- ability to test models with multiple dependent variables
- ability to model mediating variables
- ability to test coefficients across multiple groups
- ability to handle difficult data (longitudinal with autocorrelated errors, non-normal data, incomplete data)

SEM simultaneously:

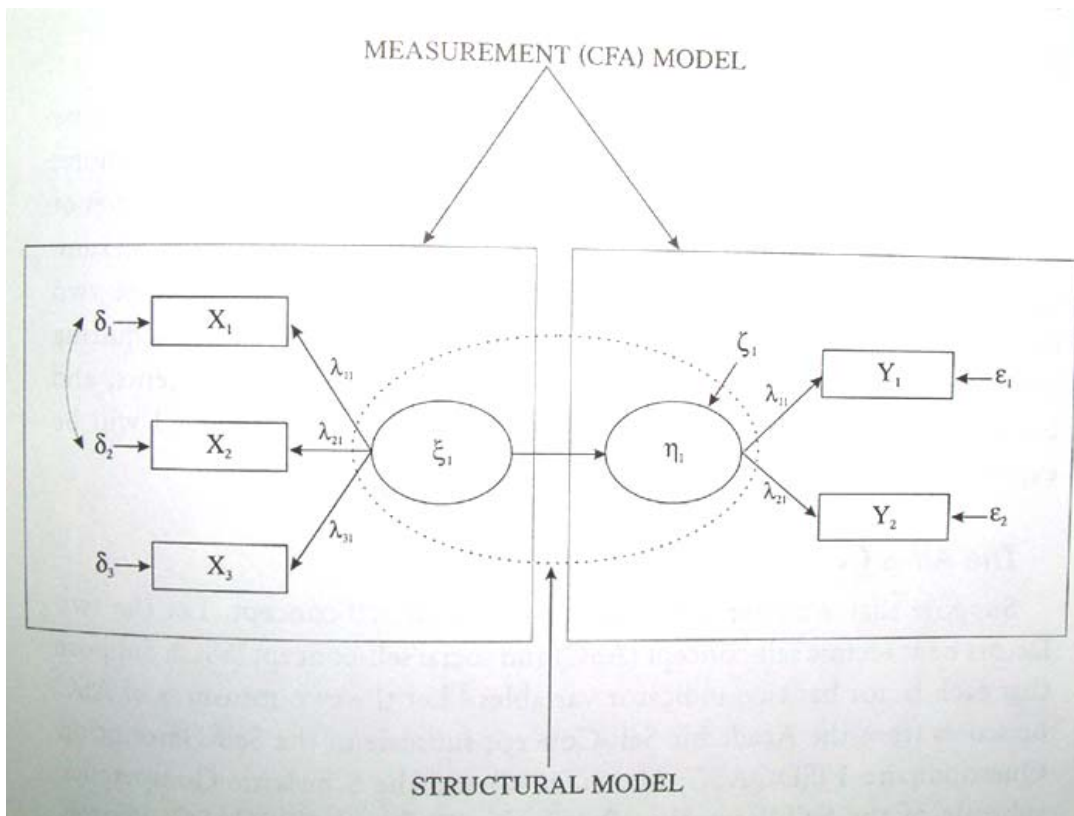
(a) models causal processes represented by a series of regression equations, and
(b) provides the ability to include unobserved (latent) variables and takes into account measurement error. In line with that, the structural equation modeling process centers around two steps:

1. Validating the measurement model -- accomplished through confirmatory factor analysis.
2. Fitting the structural model -- accomplished through path analysis with latent variables.

Sometimes, SEM can be used for only one of these two:

- (a) SEM software can be used to estimate a model in which each variable has only one indicator – i.e., to conduct path analysis
- (b) SEM software can be used to estimate a model in which each variable has multiple indicators (i.e., all variables are latent) but there are no direct effects (arrows) connecting the variables – i.e., to conduct confirmatory factor analysis

Usually, however, the term SEM refers to hybrid models with both multiple indicators for each latent variable (sometimes called factor), and directional paths specified connecting these latent variables.



1. Measurement model.

The measurement model is the part of an SEM model that deals with the latent variables and their indicators.

- Latent variables are the *unobserved variables* also called *constructs* or *factors* which are measured by their respective indicators.
- Indicators are *observed variables*, sometimes called *manifest variables* or *reference variables*, such as items in a survey instrument. Four or more indicators per latent variable is recommended, three is acceptable and common practice, two is problematic, and with one indicator, measurement error cannot be modeled (unless it is known prior to the analysis). Models using only two indicators per latent variable are more likely to fail to converge, and error estimates may be unreliable. Note: indicator variables cannot be combined arbitrarily to form latent variables. For instance, combining gender, race, or other demographic variables to form a latent variable called "background factors" would be improper because it would not represent any single underlying continuum of meaning.

A pure measurement model is a confirmatory factor analysis (CFA) model in there are straight arrows from the latent variables to their respective indicators, straight arrows from the error terms to their respective variables, but there are no direct effects (straight arrows) connecting the latent variables. In such a measurement model, we assume freely estimated covariance between each possible pair of latent variables, so we connect them with two-headed covariance arrows.

We start by specifying a model on the basis of theory. Each variable in the model is conceptualized as a latent one, even if we have to use a single indicator to measure it. Confirmatory factor analysis is used to verify that indicators seem to measure the corresponding latent variables. Note that we use common factor analysis (or principal axis factoring) rather than principal components analysis when conducting confirmatory factor analysis within SEM framework. This is important because common factor analysis assumes a separation of variance into common variance and measurement error, while principal components analysis includes all of the variance into factors it creates.

The measurement model is evaluated like any other SEM model, using goodness of fit measures (we'll discuss them in detail later). We only proceed to the structural model when we confirmed that the measurement model is valid.

2. Structural model. This step involves fitting a structural model, or multiple models if we want to compare. We can evaluate these models in terms of "model fit," which measures the extent to which the covariances predicted by the model correspond to the observed covariances in the data. If the fit is not good, we can use modification indexes to alter the model and therefore to improve fit.

Two types of variables can be identified in a structural model.

- Exogenous variables are independent variables – i.e. variables with no prior causal variables determining them (though they are usually correlated with other exogenous variables -- depicted by a double-headed arrow -- unless there is strong theoretical reason not to do so).
- Endogenous variables are dependent variables in a broad sense – these can be pure dependent variables or mediating variables (variables which are both effects of other exogenous or mediating variables, and are causes of other mediating and dependent variables).

Therefore, the structural model includes a set of exogenous and endogenous variables in the model, together with the direct effects (straight arrows) connecting them, and the disturbance terms (residual variance) for endogenous variables.

Two broad types of structural models exist -- *recursive* and *nonrecursive* models. A structural model that specifies direction of cause from one direction only is termed a recursive model; one that allows for reciprocal or feedback effects is termed a nonrecursive model. We will mostly deal with recursive models in this course, but we will address the nonrecursive ones as one of more advanced topics.

SEM is usually viewed as a confirmatory rather than exploratory procedure, using one of three approaches:

1. *Strictly confirmatory approach:* A model is generated by theory (and prior research); it is then estimated and tested using SEM goodness-of-fit tests to determine if the pattern of variances and covariances of variances in the data is consistent with the theoretical model. Because other (unexamined) models may fit

- the data just as well or even better, declaring that the model fits does not mean we confirm that it's the correct model – it just means we can't disconfirm it.
2. *Alternative models approach*: One may test two or more causal models to determine which has the best fit. There are many goodness-of-fit measures, reflecting different considerations, and usually three or four are reported by the researcher. The problem here is that it is rare to be able to find in the literature two well-developed alternative models to test – indeed, it's often not easy to find even one well-developed model.
 3. *Model development approach*: In practice, much SEM research combines confirmatory and exploratory purposes: a model is tested using SEM procedures, found to be deficient, and an alternative model is then tested based on changes suggested by SEM modification indexes. This is the most common approach found in the literature. The problem with this approach is that models confirmed in this manner are post-hoc so they may be unstable (may not fit new data, as they were based on the uniqueness of the initial dataset). Researchers may attempt to overcome this problem by using a *cross-validation* strategy -- the model is developed using one sample and then confirmed using another sample (e.g., you can split the original sample in half).

For any of these approaches, theoretical insight is key to SEM! It is especially important to realize that causal directions in SEM are inferred from the theory rather than established from the data.

Software:

LISREL, AMOS, EQS, and MPlus are four popular statistical packages for doing SEM. LISREL (LInear Structural RELations) popularized SEM in sociology and the social sciences and is still the package of reference in most articles about structural equation modeling, even though AMOS is becoming more popular because it makes it easier to specify models (because of its user-friendly graphical interface).